Centre de recherche sur les transports (C.R.T.)
Centre for Research on Transportation
CRT-2006-29

DYNAMIC TRANSPORTATION OF PATIENTS IN HOSPITALS

by

Alexandre Beaudry
Gilbert Laporte
Teresa Melo
Stefan Nickel

December 2006

Université de Montréal
C.P. 6128, succursale Centre-ville, Montréal QC H3C 3J7 Canada
Téléphone: (514) 343-7575 / Télécopieur: (514) 343-7121

crt@crt.umontreal.ca
www.crt.umontreal.ca

HEC ÉCOLE POLYTECHNIQUE MONTRÉAL Université de Montréal

2007-00277

PUBLICATION
CRT-2006-29

DYNAMIC TRANSPORTATION OF PATIENTS IN HOSPITALS

by

Alexandre Beaudry[1]
Gilbert Laporte[2]
Teresa Melo[3]
Stefan Nickel[4]

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

1. Jeppesen Ltd., 1800 McGill College Avenue, Suite 1930, Montréal, Canada H3A 3J6 (alexandre.beaudry@jeppesen.com)

2. Canada Research Chair in Distribution Management, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Canada H3T 2A7, and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada H3C 3J7 (gilbert@crt.umontreal.ca)

3. Department of Optimization, Fraunhofer Institute for Industrial Mathematics, D 67663 Kaiserslautern, Germany (melo@itwm.fraunhofer.de)

4. Fraunhofer Institute for Industrial Mathematics, and Chair of Operations Research and Logistics, Saarland University, D 66041 Saarbrücken, Germany (s.nickel@orl.uni-saarland.de)

Dépôt légal – Bibliothèque nationale du Québec
Bibliothèque nationale du Canada, 2006

Abstract

This paper analyzes and solves a patient transportation problem arising in several large hospitals. The aim is to provide an efficient and timely transport service to patients between several locations on a hospital campus. Transportation requests arrive in a dynamic fashion and the solution methodology must therefore be capable of quickly inserting new requests in the current vehicle routes. Contrary to standard dial-a-ride problems, the problem under study contains several complicating constraints which are specific to a hospital context. The paper provides a detailed description of the problem and proposes a two-phase heuristic procedure capable of handling its many features. In the first phase a simple insertion scheme is used to generate a feasible solution, which is improved in the second with a tabu search algorithm. The heuristic procedure was extensively tested on real data provided by a German hospital. Results show that the algorithm is capable of handling the dynamic aspect of the problem and of providing high quality solutions. In particular, it succeeded in reducing waiting times for patients while using fewer vehicles.

Keywords : in-house hospitals transportation, dial-a-ride, dynamic mode, tabu search.

Résumé

Cet article analyse et résout un problème de transport de patients se posant dans plusieurs grands hôpitaux. L'objectif est de fournir un service de transport efficient et ponctuel aux patients entre divers points d'un campus hospitalier. Des requêtes de transport sont formulées de façon dynamique et la méthode de résolution doit donc pouvoir les insérer dans les tournées existantes. Contrairement aux problèmes de transport à la demande standard, le problème considéré dans cet article contient plusieurs contraintes compliquées, lesquelles sont spécifiques au contexte hospitalier. L'article contient une description détaillée du problème et propose une méthode heuristique à deux phases permettant de tenir compte de ces contraintes. Dans la première phase, on utilise une méthode d'insertions simple pour construire une solution admissible, laquelle est ensuite améliorée par un algorithme de recherche avec tabous. L'heuristique a été testée en profondeur sur des données réelles fournies par un hôpital allemand. Les résultats démontrent que l'algorithme peut tenir compte de l'aspect dynamique du problème et fournir des solutions de grande qualité. En particulier, il a pu réduire les temps d'attente des patients tout en utilisant moins de véhicules.

Mots clés : transport hospitalier interne, transport à la demande, mode dynamique, recherche avec tabous.

1 Introduction

Medical diagnostic facilities, therapeutic service departments and treatment facilities are among the most visited hospital units by patients. In large hospitals, moving inpatients between health care units and service areas falls under the responsibility of a transportation department whose staff is in charge of taking patients on stretchers and beds, pushing patients in wheelchairs, and escorting ambulatory patients on foot or by ambulance to various locations within the hospital.

In-house transportation services fall under the hospital logistics area. Although these ancillary services seem simple and straightforward, they affect both core hospital services (i.e. provision of patient care) and hospital costs. For example, a late delivery of a patient to a high cost service unit, such as the operating theater suite or a magnetic resonance imaging unit, results in the underutilization of valuable resources. Furthermore, it disrupts the initially planned schedule of the department and it has, in the worst case, a domino effect: because the patient arrives after the scheduled start time, the following appointments are often delayed or the patient's appointment is given to another person thus rendering waiting time inevitable. In many hospitals, patient transportation is poorly managed, leading to patient inconvenience, disruptions, and added costs.

Despite its importance, the role played by logistics is often overlooked in health care. As reported by Landry and Philippe (2004), logistics related activities account for about 46% of a hospital's total budget. The aim of this paper is to contribute to more efficient logistics practices with respect to in-house patient transportation by proposing a new method that produces significant savings, reduces patient inconvenience and indirectly allows hospital professionals to concentrate their efforts on the quality of care they provide. Moreover, optimized patient transport processes are indispensable for reliable logistics services and a successful implementation of clinical pathways.

The core of the transportation department is the central dispatch office where new requests are received throughout the day while the vehicles are en-route servicing other rides. The dispatcher assigns incoming requests to existing vehicle routes in real-time and defines their schedule by indicating the time when each service location in the route is to be visited. Routing and scheduling decisions have a dynamic character since they are made and executed as time unfolds. Moreover, the vehicles status may deviate substantially from the conditions that were assumed when schedules were prepared, namely due to delays upon pickup and delivery of the patients. Therefore, dispatching decisions also have a temporary character since previously

constructed routes may be modified until the latest possible time before the drivers are notified of their next destinations.

Our study of the patient transportation problem is motivated by various projects carried out by the authors in several German hospitals over the past few years. We have observed that transports are frequently delayed resulting in waiting times for patients, idle times in hospital units awaiting the patients, and underutilization of staff and vehicles. The basic problem can be viewed as a dynamic dial-a-ride problem (DARP), but it is in fact considerably more complicated due to hospital-specific features. These include various order priorities, assistance of medical personnel and need for special equipment during transportation, vehicles with alternative loading modes and desired rest periods, and specific requirements for patient isolation to prevent spread of infection. These aspects significantly complicate the construction and modification of high-quality vehicle routes and schedules. Route quality is measured by two conflicting criteria, namely the minimization of fleet operating costs and the maximization of patient satisfaction. Costs are incurred for travel time and vehicle inactivity periods. Patient inconvenience is measured by lateness caused by providing service after the latest desired time. In all hospitals visited, the in-house patient transportation problem exhibited a high degree of dynamism as a result of the large majority of requests being booked at short notice. Therefore, dispatching actions are required to be very fast as time unfolds and new information becomes available.

The remainder of the paper is organized as follows. The next section provides a selective review of the literature on problems related to ours, with particular emphasis on methods proposed to solve the DARP. Section 3 presents a description of the problem under study. In Section 4 a two-phase heuristic procedure is proposed to solve the problem. Computational results are reported in Section 5 for several real-life instances from a large German hospital. Finally, conclusions and suggestions for future work are given in Section 6.

2 Literature review

The problem studied in this paper belongs to the general class of pickup and delivery problems. The DARP distinguishes itself from standard pickup and delivery problems by its focus on reducing user inconvenience. This objective is often controlled by imposing a limit on the ride time of each user (i.e., the time spent by a user in a vehicle), on the excess ride time (i.e., the difference between the actual and the minimum possible ride time of a user), and on deviations

from the desired times for pickup and delivery. Maximizing service quality is weighed against minimizing fleet operating costs, the latter being related to the number of vehicles used, total route duration and total distance traveled. The aim of the DARP is to determine a set of routes and schedules that balance these conflicting objectives. For a recent survey on the DARP the interested reader is referred to Cordeau and Laporte (2003a).

Most research on the DARP has been devoted to the static version of the problem where all user requests are known in advance and as a result, vehicle routes can be planned ahead of time. Early contributions are due to Psaraftis (1980, 1983) and Desrosiers et al. (1986) who developed dynamic programming algorithms for solving the single vehicle variant to optimality. Recently, Cordeau (2006) and Ropke et al. (2006) presented branch-and-cut algorithms for the multiple vehicle case based on the generation of several new valid inequalities. Because the DARP is $\mathcal{NP}$-hard, most of the solution procedures proposed are classical heuristics or metaheuristics that can handle large size problems. Cluster-first and route-second heuristics are common: user requests that are nearby in terms of space and time are first assigned to the same vehicle, and subsequently routing and scheduling are performed for each vehicle. Bodin and Sexton (1986), Borndörfer et al. (1997), Desrosiers et al. (1988), Dumas et al. (1989), and Wolfler Calvo and Colorni (2006) have developed different algorithms based on this approach. Various insertion strategies have also been derived, e.g. by Diana and Dessouky (2004) and Jaw et al. (1986). Toth and Vigo (1996) proposed a parallel insertion algorithm followed by an improvement step consisting of intra-route and inter-route exchanges. Parallel insertion algorithms were also developed by Fu (2002) who addressed the DARP with time-varying stochastic travel times.

Metaheuristics, in particular tabu search, are among the most successful solution methods for the static DARP. Cordeau and Laporte (2003b) have shown that a large number of variants of the DARP can be handled efficiently by the same search framework, thus rendering tabu search a powerful tool. Aldaihani and Dessouky (2003) as well as Melachrinoudis et al. (2007) presented algorithms based on tabu search. Toth and Vigo (1997) proposed a tabu thresholding algorithm for improving an initial solution obtained by the insertion procedure developed by the same authors (see Toth and Vigo (1996)). Alternative heuristics based on simulated annealing and on genetic search were designed by Baugh et al. (1998) and Bergvinsdottir et al. (2004), respectively.

While the first studies on the DARP focused on the dynamic mode (see e.g. Wilson et al. (1971) and Wilson and Colvin (1977)), this variant has received much less attention than its static counterpart. Madsen et al. (1995) have developed an insertion algorithm inspired by

the work of Jaw et al. (1986) and have applied it to a real-life dynamic problem with multiple objectives. Recently, Coslovich et al. (2006) presented a two-phase insertion algorithm for quickly deciding on the acceptance or rejection of a user request. While it makes sense to run an algorithm for a few hours in a static context, much faster response times are required in a dynamic environment. Parallel computing is therefore a natural way of reducing computing time. Attanasio et al. (2004) proposed several parallel implementations of the tabu search procedure of Cordeau and Laporte (2003b).

3 Problem description

This section describes the patient transportation problem. We visited several large German hospitals having at least 1000 beds, and gathered information on their patient transportation services. Our aim was to define a problem that captures the characteristics of different transportation systems. We believe that our problem description is generic enough to support a broad range of practices that are prevalent in hospitals not only in Germany but also in other parts of the world.

3.1 Transportation requests

As in the standard DARP, a request for transportation consists of an origin location where the patient is picked up, a destination location where the patient is dropped-off, a desired time instant either for pickup or delivery, and the mode of transportation required (e.g. a wheelchair or a stretcher). Furthermore, to limit patient inconvenience a maximum ride time is specified.

In a hospital context, additional data are provided upon order entry. Each request is booked with a given priority indicating its degree of urgency. This information is conveyed by a maximum deviation allowed from the desired time instant for service, and is used to define soft time windows both at the pickup and drop-off locations as will be described below. The medical condition of the patient may not allow sharing the ambulance with other patients during the transport. This is often the case when it is necessary to isolate the patient to prevent spread of infection. Also, additional equipment, such as an oxygen tank or a portable electrocardiogram (ECG) monitoring device, may be required. Finally, the patient may need the assistance of medical or nursing staff during the transport. When the patient and the accompanying person are in different hospital units, the medical escort is picked up before the patient. The medical escort may stay at the delivery location of the patient or be brought to another location, usually

the original point. All accompanying persons travel seated when an ambulance is used. Since staff time is costly, a transportation request needing nursing or medical assistance is treated as a *chain of requests* that cannot be interrupted by servicing other requests. In this case, a chain may include up to four locations. Table 1 summarizes the notation associated with a request.

$o(i), d(i)$	origin, resp. destination, location of request i
r_i^m	number of persons requiring transportation mode m, e.g. $m = S$ for seated transport, $m = L$ for transport with bed, $m = W$ for transport with wheechair
q_i^e	$= 1$ if equipment of type e is required for servicing request i, e.g. $e =$ oxygen tank, $e =$ portable ECG device, 0 otherwise
w_i	$= 1$ if patient isolation is required (i.e. no ride-sharing), 0 otherwise
y_i	time required to disinfect a vehicle due to patient isolation
t_i^p, t_i^d	earliest, resp. latest, desired time instant for pickup, resp. delivery, of request i
Δt_i	maximum deviation allowed from the desired service time instant of request i
MRT_i	maximum acceptable ride time of request i
EPT_i, LPT_i	earliest, resp. latest, pickup time at the origin location of request i
EDT_i, LDT_i	earliest, resp. latest, drop-off time at the destination location of request i
$DRT_{o(i),d(i)}$	direct ride time from $o(i)$ to $d(i)$ including dwell times

Table 1: Notation associated with a transportation request

The shortest travel time between any two stops is assumed to be known, and accounts for driving on the campus grounds and walking inside the buildings. Usually, direct ride times are not symmetric due to one-way streets. Dwell times are added to the direct ride time and include not only the time needed to load and unload the patient into and from the ambulance, but also additional time spent at the origin and destination locations. At nursing wards, the transporter transfers the patient from the bed to a stretcher or a wheelchair, and collects the patient's medical record from the nurses' office. At service departments, the transporter informs a nurse that he is picking up or delivering a patient. We assume that the time required to perform these tasks does not significantly depend on the transportation mode since the staff transporting patients is highly experienced.

Each request is submitted with a desired time either for pickup at the origin or for delivery at the destination. Since travel times are to some extent uncertain, time windows are defined by combining this information with a prespecified tolerance that reflects the priority of the request. When the earliest desired pickup time t_i^p for request i is specified, the origin $o(i)$ is called *critical location* and the time windows are defined as follows: $EPT_i = t_i^p$, $LPT_i = t_i^p + \Delta t_i$

and $EDT_i = EPT_i + DRT_{o(i),d(i)}$, $LDT_i = LPT_i + DRT_{o(i),d(i)}$. For appointments at service units, each corresponding transportation request is usually booked with a desired latest delivery time t_i^d. Stop $d(i)$ is thus the *critical location* and the time windows are set as follows: $LDT_i = t_i^d$, $EDT_i = t_i^d - \Delta t_i$ and $LPT_i = LDT_i - DRT_{o(i),d(i)}$, $EPT_i = EDT_i - DRT_{o(i),d(i)}$. These settings are in line with the policy that a patient cannot be left unattended upon delivery. A member of the nursing staff must be notified of the arrival of the patient. The notion of critical location was introduced by Cordeau and Laporte (2003b) and will play an important role in our solution approach. For a chain of requests, the critical location corresponds to the first stop in the chain regardless of the location with the desired service time. Moreover, the time windows at the locations not associated with the pickup and delivery of the patient are easily obtained by taking the corresponding direct ride times.

It is assumed that each pickup location $o(i)$ cannot be serviced before EPT_i. Service after LPT_i is, however, allowed. Due to the manner in which time windows are set, an early arrival at a delivery location is not possible but a late arrival may occur. Non-compliance with the time windows at the critical locations leads to penalties (see Section 3.3 for further details). At first sight the time windows may seem too strict since $DRT_{o(i),d(i)} \leq MRT_i$. However, only deviations at the critical locations are evaluated and as a result, penalties at non-critical locations are not considered.

3.2 Vehicle fleet

Depending on the hospital layout (i.e. one central building as opposed to a campus with several buildings), intrahospital transports are provided by specially equipped vehicles, typically ambulances, or ambulatory staff. In some hospitals both transportation modes are available. For convenience, the term *vehicle* will be used to denote any transportation resource. Table 2 summarizes the notation associated with a vehicle.

Transportation requests are serviced by a heterogeneous fleet of vehicles housed in different depots. Each vehicle is available during a given period of the day with one or more scheduled service interruptions with fixed duration. A desired start time is assigned to each break period and a positive or negative deviation is allowed as long as it does not exceed a prespecified tolerance. Work breaks refer to rest periods of drivers as well as regular vehicle maintenance activities, and are taken at the vehicle's origin depot. This policy, which is followed by many hospitals, provides flexibility for building vehicle routes. Each vehicle carries special equipment

$o(v), d(v)$	origin, resp. destination, depot of vehicle v, typically $o(v) = d(v)$
$R^{\ell}_{v,m}$	capacity of vehicle v for transportation mode m with loading alternative ℓ
Q^{e}_{v}	number of pieces of equipment of type e available in vehicle v
t^{p}_{v}, t^{d}_{v}	earliest departure, resp. latest arrival, time from the origin, resp. at the destination depot of vehicle v
BT^{k}_{v}	desired start time of the kth work break of vehicle v
BD^{k}_{v}	duration of the kth work break of vehicle v
ΔB^{k}_{v}	maximum allowed deviation from the desired begin of the kth break of vehicle v

Table 2: Notation associated with a vehicle

such as portable ECG devices and oxygen tanks, and has alternative ways of loading different transportation modes. The latter are modeled as disjunctive multi-dimensional capacities. For example, the vehicle may transport up to one bed, one wheelchair and one seated ambulatory patient at the same time. Alternatively, up to three patients can be transported seated in wheelchairs. Upon completion of the transport of a patient on isolation the vehicle returns to its origin depot for disinfection. This leads to a chain of requests consisting of three stops: the pickup and delivery locations for the patient, followed by the vehicle's depot. The width of the time window at the depot is set by the duration of the cleaning activities performed by the driver. Moreover, the earliest arrival time at that location equals the latest departure time from the drop-off point of the patient plus the direct ride time. Finally, vehicles with patients on board are only allowed to wait at a pickup location if they arrive before the earliest desired pickup time. On the other hand, a policy in place at several German hospitals is for an empty vehicle to return to its depot to avoid idling at a pickup location.

3.3 Constraints and objective

The aim of the patient transportation problem is to design vehicle routes and schedules for intrahospital requests in an effective way, subject to (i) visiting constraints: each pickup and delivery location is visited only once; (ii) depot constraints: the route of each vehicle starts and ends at the corresponding depots; (iii) capacity constraints: vehicle capacities and pieces of equipment cannot be exceeded; (iv) pairing constraints: the pickup and delivery locations of each request must be visited by the same vehicle; (v) precedence constraints: each patient must be picked up before being dropped off; (vi) patient inconvenience constraints: the time elapsed between the pickup and delivery of each patient cannot exceed a prespecified maximum

ride time; (vii) resource constraints: the service periods of the vehicles must be respected and deviations from the desired begin of break periods cannot exceed a given tolerance; (viii) hospital service constraints: requests for single transportation must be fulfilled and the sequence of stops in each chain of requests must be satisfied without intermediate visits to other locations.

Route effectiveness is measured by a good tradeoff between two conflicting objectives. On the one hand, the hospital wishes to minimize direct and indirect operating costs. On the other hand, patient inconvenience is to be avoided. Travel time is one of the factors having a direct impact on the operating costs of the transportation department. Violation of the time windows causes not only inconvenience to the patient but also impacts negatively on hospital costs. In fact, a late pickup at the origin location leads to waiting time for the patient, while an early arrival generates idle time for the vehicle crew which is reflected in increased direct costs. Regarding the destination location, an early arrival is not possible due to the way the time window is set. However, late delivery adds not only to the patient inconvenience but also to the hospital indirect costs. This is particularly true when the delivery location is a service department (e.g. an operation theater) since it results in underutilization of equipment and staff. In addition, late delivery causes disruption of the planned schedule at the service unit, and the patient may end up having to wait a considerable amount of time for his delayed appointment.

Let $\mathcal{T}_v = \{n_1, \ldots, n_q\}$ denote the route of vehicle v which consists of a feasible sequence of service points n_k (for pickup and delivery). Table 3 describes the notation associated with the scheduled times, earliness and lateness at each stop. As mentioned before, time window devi-

$\mathcal{T}_v$	route of vehicle v consisting of stops $n_1, n_2, \ldots, n_q$ with $n_1 = o(v)$ and $n_q = d(v)$
$A^v_{n_k}$	projected arrival time of vehicle v at stop n_k
$D^v_{n_k}$	projected departure time of vehicle v from stop n_k
$E^v_{n_k}$	earliness of vehicle v upon arrival at stop n_k
$L^v_{n_k}$	lateness of vehicle v upon arrival at stop n_k
$C^{v,m}_{n_k}$	load of vehicle v for transportation mode m upon departure from stop n_k
$S^{v,e}_{n_k}$	pieces of equipment of type e in use in vehicle v upon departure from stop n_k
$f(E^v_{n_k})$	penalty for early arrival at stop n_k
$g(L^v_{n_k})$	penalty for late arrival at stop n_k

Table 3: Notation associated with a vehicle route

ations at critical locations are penalized in an attempt to keep costs and patient inconvenience low. Let i_k denote the transport request associated with stop n_k. A linear penalty function

is considered for early arrivals at pickup locations that are also critical: $f(E^v_{n_k}) = p^e_{i_k} E^v_{n_k}$ with $p^e_{i_k}$ denoting the penalty factor for earliness which depends on the urgency of the request. A piecewise linear function is applied to lateness at critical locations: $g(L^v_{n_k}) = p^1_{i_k} L^v_{n_k}$ if $0 \leq L^v_{n_k} \leq \Delta t_{i_k}$, and $g(L^v_{n_k}) = p^2_{i_k} L^v_{n_k}$ if $L^v_{n_k} > \Delta t_{i_k}$, with $p^1_{i_k}$ and $p^2_{i_k}$ denoting penalty terms such that $p^1_{i_k} < p^2_{i_k}$. Observe that the breakpoint Δt_{i_k} reflects the priority of the request and thus, the left slope ($p^1_{i_k}$) is lower than the right slope ($p^2_{i_k}$). In other words, lateness exceeding the maximum desired variation Δt_{i_k} is less tolerated and therefore generates a higher penalty. At non-critical stops n_k we set $f(E^v_{n_k}) = g(L^v_{n_k}) = 0$. Both penalty functions are in line with the linear inconvenience functions proposed by Sexton and Choi (1986), and Toth and Vigo (1997). Bergvinsdottir et al. (2004) also use a piecewise linear function with increasing slopes depending on the arrival time at each destination location. The cost function of our problem consists of minimizing a weighted sum of three criteria: total travel time, total lateness, and total earliness. Thus, the resulting function $c(x)$, with x denoting a solution, is

$$c(x) = \omega_1 \sum_v \left(A^v_{n_q} - D^v_{n_1} - \sum_{k=1}^{q} E^v_{n_k} \right) + \omega_2 \sum_v \sum_{k=1}^{q} g(L^v_{n_k}) + \omega_3 \sum_v \sum_{k=1}^{q} f(E^v_{n_k}) \quad (1)$$

where ω_j are non-negative parameters, $j = 1, 2, 3$. These weights reflect the relative importance of the above criteria.

3.4 Dynamic events

In the German hospitals we visited, the percentage of transportation requests booked one day in advance is rather small (less than 10%). This is due to the fact that ward rounds with physicians take place early in the morning and afterwards an increase in transport bookings to diagnostic and therapeutic facilities is observed. Hence, the patient transportation problem arises in a dynamic context as data are gradually revealed during the day. In such an environment, a planned vehicle route is a sequence of requests that have been assigned to the vehicle and scheduled but have not been serviced yet. Typical events that trigger modifications of previously planned routes include (i) patient-related events: arrival of new requests for transportation, changes in the data of not yet serviced requests as a result of erroneous information, and cancellation of requests; (ii) vehicle-related events: arrival of a vehicle at a service location, vehicle breakdown, and unscheduled rest periods. As a rule, dispatching actions are required to be very fast upon disclosure of new information.

At each service location, the driver notifies the transportation department of the end of

service and is assigned a new destination. At any point in time drivers do not have knowledge of their current planned route entirely, which allows changes in the routes if necessary. It is assumed that the exact position of each moving vehicle is not known to the dispatch office and therefore, once a vehicle is driving to the next stop, it cannot be diverted away from its current destination.

Due to the complexity and combinatorial nature of the problem, we choose not to present a mathematical formulation since it would not provide useful insight into the structure of good feasible solutions. In fact, finding a feasible solution for the DARP is itself $\mathcal{NP}$-complete in the strong sense, as shown by Savelsbergh (1985). Moreover, a dynamic environment calls for a tradeoff between solution quality and response time.

4 Algorithm

We propose a two-phase heuristic to be run between the occurrence of new events. The aim of the first phase is to insert a new request into an existing route by following a simple, yet powerful, insertion scheme. It makes use of the notion of spatial and temporal proximity between the new request and all other requests already scheduled but not yet serviced. A list of tentative neighbors is created and used as reference positions for inserting the new request. The second phase is a tabu search algorithm that considers intermediate infeasible solutions during the search process in an attempt to obtain a better solution. Two different types of moves are implemented: the first removes a request from its current route and reinserts it in a different route; the second consists in rearranging the sequence of not yet serviced requests in their assigned route.

In view of the pairing constraints and the fact that vehicle diversion is not allowed, at any point in time some of the unvisited stops in a vehicle route cannot be removed and assigned to another vehicle. Such stops relate to so-called *pre-assigned* requests. They include the current destination point of the vehicle and the delivery locations of patients on board at that time.

4.1 Route initialization

The algorithm starts by initializing the routes of the vehicles according to their service periods which are regulated by the work schedules of the transportation staff. The assignment of crews to vehicles is decided beforehand by the transportation department. The first stop of each vehicle route is the origin depot of the vehicle, while the last stop is its destination depot.

Transportation requests can only be inserted between these two stops. The time windows at the depots have zero width, and correspond to the start (t_v^p) and end times (t_v^d) of the crew's work shift. In some hospitals, deviations from scheduled working hours are allowed provided they do not exceed some tolerance. The latter can then be used to set the width of the time windows at the depots. Work breaks planned beforehand are modeled as special transportation requests and inserted in the initial route in chronological order. Each kth service break of vehicle v has the same origin and destination location, usually $o(v)$. The "travel" time between these locations is equal to the break duration, which also limits the maximum ride time. The time window at the "pickup" point, which is the critical location, is defined by $EPT_i = BT_v^k - \Delta B_v^k$ and $LPT_i = BT_v^k + \Delta B_v^k$. "Pickup" after LPT_i incurs a very large penalty. At the "delivery" location the time window is set by $EDT_i = EPT_i + BD_v^k$ and $LDT_i = LPT_i + BD_v^k$. This special request has no demand or equipment requirements (i.e. $r_i^m = 0$ for every transportation mode m, $q_i^e = 0$ for every piece of equipment e). To guarantee that the vehicle is empty during the break period, "patient isolation" is imposed by setting $w_i = 1$.

4.2 Neighborhoods

Neighborhood search heuristics have been applied with success to a variety of pickup and delivery problems, see e.g. Bent and Hentenryck (2006), Gendreau et al. (1998a), Nanry and Barnes (2000), and Ropke and Pisinger (2006). To group together similar requests, the level of attraction between two requests is measured by the proximity between their pickup and delivery locations. As observed by Gendreau et al. (1998b), it does not suffice to define neighborhoods using only distances or travel times, since two locations close to each other may have time windows that are far apart or even incompatible. Hence, both the spatial and time dimensions have to be considered in the characterization of a neighborhood. In view of the definition of the time windows and the evaluation of service quality in our problem, the proximity between two requests i and j is determined by the attraction level of their critical locations. Therefore, the neighborhood $N_p(i)$ of an unassigned request i is the set of p critical stops that are closest to the critical location of i. Closeness is measured by the spatial proximity of p_1 stops and by the temporal proximity of p_2 stops, where p_1 and p_2 are user-defined parameters such that $p_1 + p_2 = p$. Each subset $N_{p_k}(i)$ is further decomposed into successors and predecessors of the critical location of i, denoted resp. $N_{p_k}^+(i)$ and $N_{p_k}^-(i)$ for $k = 1, 2$.

Consider two requests i and j. There are six different ways to form a path with these

requests:

$$o(i) \rightarrow o(j) \rightarrow d(i) \rightarrow d(j) \tag{2}$$

$$o(i) \rightarrow o(j) \rightarrow d(j) \rightarrow d(i) \tag{3}$$

$$o(j) \rightarrow o(i) \rightarrow d(i) \rightarrow d(j) \tag{4}$$

$$o(j) \rightarrow o(i) \rightarrow d(j) \rightarrow d(i) \tag{5}$$

$$o(i) \rightarrow d(i) \rightarrow o(j) \rightarrow d(j) \tag{6}$$

$$o(j) \rightarrow d(j) \rightarrow o(i) \rightarrow d(i). \tag{7}$$

To measure the spatial proximity between i and j, the shortest path that links their critical locations directly is determined. Let d_{ij}, resp. d_{ji}, denote the proximity when the critical location of request i is serviced before, resp. after, the critical location of request j. If $o(i)$ and $o(j)$ are critical then

$$\begin{aligned} d_{ij} &= DRT_{o(i),o(j)} + \min\left\{DRT_{o(j),d(i)} + DRT_{d(i),d(j)}, DRT_{o(j),d(j)} + DRT_{d(j),d(i)}\right\}, \\ d_{ji} &= DRT_{o(j),o(i)} + \min\left\{DRT_{o(i),d(i)} + DRT_{d(i),d(j)}, DRT_{o(i),d(j)} + DRT_{d(j),d(i)}\right\}. \end{aligned}$$

If $d(i)$ and $d(j)$ are critical then d_{ij} corresponds to the shortest duration between paths (2) and (4). In contrast, d_{ji} is obtained by comparing path (3) with path (5). If $o(i)$ and $d(j)$ are critical then d_{ij} is the length of path (5), while d_{ji} is the length of path (7). Finally, if $d(i)$ and $o(j)$ are critical, then d_{ij} corresponds to the duration of path (6) and d_{ji} is the duration of path (2).

Due the the dynamic nature of our problem, request j may be pre-assigned, meaning that either $o(j)$ was already visited by a particular vehicle v or $o(j)$ is the immediate stop of that vehicle. In the first case, if $d(j)$ is the next destination of the vehicle, then only the path $n \rightarrow d(j) \rightarrow o(i) \rightarrow d(i)$ is of interest, with n denoting the predecessor of $d(j)$ in the planned route of vehicle v. Furthermore, the proximity only has to be calculated if $d(j)$ and $o(i)$ are both critical. When $o(j)$ has been serviced but $d(j)$ is *not* the immediate destination of the vehicle, paths (4), (5) and (7) are considered with $o(j)$ replaced by the current stop n of the vehicle. Clearly, these paths only contribute to the spatial proximity of requests i and j when location $d(j)$ is critical. Finally, in the case $o(j)$ is the immediate stop of the vehicle then paths (4), (5) and (7) are analyzed following the same rules as for the case of request j not being pre-assigned.

The temporal proximity between two requests i and j is based on the evaluation of the compatibility of the time windows of their critical locations and also depends on their position

in paths (2)–(7). The derivation of this attractiveness measure is rather technical, and is therefore presented in the appendix.

4.3 Insertion phase

Since the transportation department continuously receives new requests, a decision must be quickly taken about the allocation of a new request to a particular vehicle and its scheduling within the vehicle's planned route. This is the role of the insertion heuristic. Especially in the morning peak hours, consecutive events may occur within a few seconds and as a result, a new solution must be computationally inexpensive. Because response time is a crucial issue, we have opted in this phase for a simple procedure that preserves feasibility at all times. If additional time is available, the post-optimization phase is run with a relaxation of this requirement in an attempt to find solutions of superior quality by exploring infeasible regions.

The insertion heuristic is run whenever a request needs to be scheduled. A rolling horizon of length ΔM is used, so that only new requests with a latest desired delivery time in the next ΔM time units are considered. This is a prevalent practice in dynamic pickup and delivery problems which avoids investing computational effort to handle long-term events. The processing of a (new) request, say i, starts by constructing a list of neighboring stops among those in the planned vehicle routes. This entails the evaluation of the spatial and temporal proximity measures described in Section 4.2. Each neighboring stop refers to a tentative successor or predecessor of the critical location of i, and therefore, suggests an insertion position in a particular vehicle route. Holding the critical location in that position, and maintaining the ordering of the stops already in the route, all possible positions for the corresponding non-critical stop are examined. Due to the precedence constraints, there are only a limited number of possible positions. Among these, only feasible positions both for request i as well as for all other trips already assigned to that vehicle are retained. Moreover, an insertion between intermediate stops of a chain of requests is not allowed. Handling critical locations in this way helps reduce the number of computations. This procedure is applied to every neighbor in $N_p(i)$. The origin and destination locations of request i are then inserted in the overall best feasible positions found. The quality of each pair of feasible insertion positions in a given vehicle route is measured by the term in the cost function (1) referring to that route.

Just before the morning shift starts, the route initialization procedure described in Section 4.1 is applied. At that moment some of the requests received the day before and during

the night await to be processed. Requests demanding service within the next ΔM time units are then considered. These requests are sorted by non-increasing value of Δt_i which is a priority indicator. Ties are broken by taking the requests in increasing order of their earliest pickup times. Each request is then processed in sequence by the insertion heuristic until the list is exhausted.

4.4 Improvement phase

Tabu search (TS) has proved to be a highly effective metaheuristic for the vehicle routing problem and many of its variants. For the DARP, recent examples include Aldaihani and Dessouky (2003), Attanasio et al. (2004), and Cordeau and Laporte (2003b). The latter reference provides the framework for our TS algorithm.

The exploration of the solution space is guided by two types of moves: the first removes the pickup and delivery stops of an unserviced request from its current route and reinserts them in a different route (*inter-route moves*). The second consists of rearranging the sequence, and thereby the schedule, of not yet serviced stops within their assigned route (*intra-route moves*). These moves define the neighborhood structures used by the algorithm.

The TS algorithm starts from the solution x_0 obtained by the insertion heuristic and chooses at each iteration ℓ the best non-tabu solution in the neighbourhood of x_ℓ among the inter-route moves. This step includes a diversification strategy that forces the search to explore less visited regions of the solution space. To intensify the search in promising regions, intra-route moves are performed every κ iterations and whenever a new best feasible solution is identified. The TS algorithm stops either when the maximum number of iterations ℓ_{max} is attained or when a new event occurs that has to be processed. The larger the time elapsed between two consecutive events is, the greater the expected improvement potential will be.

4.4.1 Inter-route moves

Each iteration starts by selecting an outstanding request, without a pre-assigned status, for removal from its current vehicle route according to a given criterion, e.g. the request having the largest lateness. In Section 5.2 different criteria will be described. Let (i, v) denote the request-vehicle pair, also called an *attribute*, in the current solution. Next, the vehicles having the necessary characteristics to satisfy request i are considered (i.e. enough capacity: $R^{\ell}_{v,m} \geq r^m_i$, appropriate equipment: $Q^e_v \geq q^e_i$, and suitable working hours: $EPT_i,\ LDT_i \in [t^p_v, t^d_v]$).

Following the scheme described in Section 4.2, p neighbors of the critical location of request i are identified in the routes of the pre-selected vehicles. For each neighbor, tentative insertion positions for the pickup and drop-off points of request i are evaluated in a similar way as in the insertion heuristic. However, contrary to that method, a wider exploration of the search space is now permitted by accepting positions yielding infeasible routes. This approach is used in an attempt to find good feasible solutions by means of a series of route modifications that violate the capacity, equipment and maximum ride time constraints. An intermediate solution containing infeasible routes is evaluated by the penalized cost function

$$c'(x) = c(x) + \alpha\, R'(x) + \beta\, Q'(x) + \gamma\, M'(x), \tag{8}$$

where $c(x)$ is given by (1), $R'(x)$, $Q'(x)$ and $M'(x)$ measure the total excess capacity, total excess equipment and total excess ride time of all routes, respectively, and α, β and γ are positive penalty factors. The latter are initially set to 1 and will be adjusted during the search. If a solution is infeasible with respect to a constraint, the associated penalty factor is multiplied by $1+\mu$, otherwise it is divided by $1+\mu$, with μ being a user-defined parameter. Cordeau and Laporte (2003b) showed that this relaxation mechanism adds diversity to the search through the exploration of infeasible regions and fits well with simple neighborhood structures such as the inter-route moves described above. Among all the moves considered, the move that minimizes (8) is selected. To avoid cycling, a request transferred from its current route $\mathcal{T}_v$ to another route $\mathcal{T}_{v'}$ cannot be reinserted in route $\mathcal{T}_v$, and is therefore *tabu*, for θ iterations. Through an aspiration criterion, the tabu status can be revoked if such a reverse move would result in a new solution with lower cost among all known solutions containing attribute (i, v). The new solution has the same set of attributes as the previous one except that the pair (i, v) is replaced by the new attribute (i, v').

4.4.2 Diversification mechanism

As in most TS implementations, we also penalize frequently performed moves. Let ρ_{iv} denote the number of times that attribute (i, v) has been added to a solution during the search (i.e. request i has been assigned to vehicle v). If (i, v) is the attribute added to solution x_ℓ in the ℓth iteration to obtain solution x and $c'(x) > c'(x_\ell)$, then the penalty factor $\lambda\, c(x)\, \sqrt{n\, m}\, \rho_{iv}/\ell$ is added to $c'(x)$. The term $\sqrt{n\, m}$ measures the number of possible attributes, which depends on the number n of requests that can be moved and on the number m of available vehicles. The user-defined parameter λ controls the intensity of the diversification. By multiplying the cost of

solution x by an instance-dependent factor, namely, $c(x)\sqrt{n\,m}$, the more attributes there are, the higher a frequently added attribute is penalized. This mechanism was first implemented by Taillard (1993).

4.5 Route adjustments

In a dynamic environment, the dispatching process evolves during the day. This calls for suitable mechanisms to handle new events, keep track of the planned routes and implement the necessary changes. Thus, the proposed algorithm is embedded within a framework that will allow modifying and updating previously planned routes as time unfolds. An incoming request triggers the two-phase optimization procedure. This also occurs when the desired service time of an outstanding request is postponed by the user. In this case, the request is first removed from its current route and the remaining stops are joined while maintaining their ordering in the route. Upon notification of the end of service at a location, the service point (pickup or delivery) is removed from its route, the actual time is saved, the scheduled times of the remaining stops in the route are updated, and the driver is notified of the next destination. The improvement phase is run to handle significant changes in the new projected service times. For example, if the reported end of service occurs later than expected, this will allow moving one or more outstanding requests to other vehicles to try to avoid lateness. Similar actions are performed for an arrival event, which is reported later than expected. The cancellation of a request also triggers the second phase of the algorithm after having removed the service points from their route.

Our algorithm employs a drive-first waiting strategy for scheduling the trips, meaning that a vehicle drives from a location at the earliest departure time. Therefore, the scheduled arrival time at a service location is the earliest possible time. To prevent an empty vehicle from idling at a pickup location, several German hospitals deviate from this policy by deliberately sending the vehicle back to its home depot. Hence, at the end of the optimization run, if stop n_k is a pickup location in route $\mathcal{T}_v$, $E^v_{n_k} > 0$, $C^{v,m}_{n_{k-1}} = 0$ for every transportation mode m (i.e. the vehicle is empty), and $n_{k-1} \neq o(v)$, the possibility of inserting the depot before stop n_k is analyzed. This is the case when condition

$$D^v_{n_{k-1}} + DRT_{n_{k-1},o(v)} + \Delta W + DRT_{o(v),n_k} \leq LPT_{i_k}$$

holds with i_k denoting the request associated with stop n_k, and ΔW being a user-defined minimum length of stay at the depot. The insertion of $o(v)$ is revoked each time the heuristic

procedure is run unless the depot is already the immediate destination of the vehicle.

5 Computational experiments

We now report on the results obtained by applying the proposed algorithm to real data from a large German hospital. The aim is to compare different variants of the algorithm for handling new requests under various operational conditions. In the following, the patient transportation service of the hospital is introduced and the operating scenarios considered for the computational experiments are described. The parameter settings are presented, followed by the numerical results.

5.1 Hospital data

The hospital considered is an academic health care institution, located in south-west Germany, with 1400 beds and a staff of approximately 4000. Figure 1 shows the layout of the hospital campus which spreads over a large area with more than 100 buildings and includes a road network of 15 km. The hospital operates a fleet of 11 ambulances housed in two depots (buildings 79 and 88) and transports about 250 patients daily. Each vehicle is operated by a team of two people who work day shifts as described in Table 4. Outside normal work hours transports are subcontracted to a private ambulance service. During each work shift a 30-minute rest period with a desired start time is scheduled. The work break can begin at most 10 minutes before or after the specified time. All vehicles move at a constant speed of 10 km/h which accounts for driving on campus, walking inside the buildings, as well as loading and unloading operations. Moreover, vehicles can only return to their home depots during the day if their sojourn time is at least $\Delta W = 10$ minutes.

All ambulances, except vehicle 3, have the same capacity and are equipped with one oxygen tank. Vehicle 3 cannot transport beds and cannot be used when patient isolation is requested. Table 5 describes the disjunctive multi-dimensional capacities by vehicle type. Requests for transportation are classified as *normal* or *urgent* requests. In the first case, a patient can spend at most 30 minutes in an ambulance and the time windows have a 15-minute width. For urgent requests a maximum ride time of 21 minutes is imposed and the width of the time windows is five minutes. As shown in Table 6, the daily fraction of urgent requests is less than 2%.

Concerning the optimization criteria, the hospital gives higher priority to patient convenience over travel time and early arrival at pickup locations. Therefore, the weighting parameters in

Figure 1: Hospital campus

the objective function are $\omega_1 = 0.3$, $\omega_2 = 0.6$ and $\omega_3 = 0.2$. This choice is also reflected in the penalty factors for earliness and lateness at critical locations, namely $p^e_{i_k} = p^1_{i_k} = 1$ and $p^2_{i_k} = 3$. This means that pickups or deliveries regarding normal, resp. urgent, requests incur a small penalty as long as lateness is below 15, resp. five, minutes. Lateness beyond these tolerance values is penalized by a factor of three. On the other hand, the penalty incurred to the arrival of a vehicle before the earliest desired time for pickup is simply the number of minutes early.

Requests for travel are booked online over the intranet by hospital units and handled on a first-come-first-served basis by the transportation department. Due to the possibility of ride-

Vehicle			Service		Work break	
Number	Type	Depot	Start	End	Start	End
1	1	1	07:48	16:00	12:00	12:30
2	1	1	07:48	16:00	12:00	12:30
3	2	1	08:30	16:42	13:00	13:30
4	1	1	07:00	15:12	12:00	12:30
5	1	1	08:00	16:12	12:30	13:00
6	1	1	07:30	15:42	12:00	12:30
7	1	1	08:00	16:12	12:30	13:00
8	1	2	08:30	16:42	13:00	13:30
9	1	1	07:00	14:12	11:30	12:00
10	1	1	08:00	16:12	12:30	13:00
11	1	1	07:00	14:12	11:30	12:00

Table 4: Vehicle availability

sharing, multiple requests are assigned simultaneously to a vehicle by the dispatcher. Communication between the dispatcher and the vehicles is maintained via hand-held radio transceivers. Since routing and scheduling decisions are made by the vehicle crew but the dispatcher is not notified, the central dispatch office is only called by the vehicle upon completion of all jobs, that is, once the vehicle is empty. The sole exception to this rule is in the case of an urgent request, when the dispatcher contacts several vehicles over the radio to be informed of their current position and workload. Due to partial availability of information, it is not possible to automatically create management records of the completed journeys by the end of the day, and

Vehicle	Loading	Capacity		
type	alternative	Seats	Beds	Wheelchairs
1	1	4	2	1
1	2	5	2	0
1	3	1	0	0
1	4	0	1	0
1	5	0	0	1
2	1	6	0	1
2	2	7	0	0

Table 5: Alternative loading modes per vehicle type

as a result, the quality of the services provided cannot be fully assessed. However, patients and hospital staff complain that transports are frequently delayed, which leads to undesired patient waiting times, and underutilization of staff and equipment resources at the service units awaiting the patients. Furthermore, sometimes transportation staff carries patients without explicit notification from the dispatcher. This results in an empty trip when the same request is assigned to another vehicle.

Day	Requests		% Requests booked				
	Total	% Urgent	Day before	Same day with booking time			
				after EPT	before EPT (minutes)		
					]0, 10]	]10, 30]	> 30
1	265	0.8	5.3	1.9	43.8	25.3	23.8
2	248	0.0	2.0	3.2	53.6	18.5	22.6
3	216	0.0	3.2	0.5	49.1	22.2	25.0
4	291	1.0	4.5	0.7	50.2	19.6	25.1
5	259	0.0	4.6	0.8	45.6	18.9	30.1
6	207	0.5	4.3	1.0	49.8	15.5	29.5
7	226	0.0	4.0	0.4	47.3	23.0	25.2
8	302	1.0	6.0	0.3	51.3	17.5	24.8
9	182	0.0	2.7	0.0	50.0	23.6	23.6
10	213	0.9	4.2	0.5	49.8	16.9	28.6
11	258	0.8	3.1	0.8	50.8	16.3	29.1
12	224	1.3	2.2	0.9	49.6	17.0	30.4
13	239	1.3	6.3	0.0	43.9	24.7	25.1
14	286	0.0	1.4	0.0	49.7	18.9	30.1
15	242	0.0	6.2	0.4	43.4	24.0	26.0
16	195	0.0	2.6	0.0	53.3	17.4	26.7
17	241	1.2	2.9	0.8	53.9	16.6	25.7
18	210	0.5	2.4	0.0	48.6	20.5	28.6
19	284	0.4	2.8	0.4	53.5	19.7	23.6
20	237	0.0	4.6	0.0	48.5	24.1	22.8
Avg.	240	0.5	3.8	0.6	49.3	20.0	26.3

Table 6: Booking times compared with the desired earliest pickup times (EPT) of requests

The hospital supplied data regarding all requests serviced over a 20-day period. The data include the characteristics of each request along with the corresponding booking time. Table 6 presents the daily number of requests serviced and the proportion of urgent requests. In addition, it reports the time lag between the booking time of a transport request and its desired earliest

pickup time (EPT). Advanced notification of requests is rather seldom with only 3.8% of the requests being ordered the day before. In other words, the data exhibit a strong degree of dynamism, being on average $\delta = 0.96$, with δ equal to the number of requests with occurrence time on the same day of service divided by the total number of requests. The majority of transport needs are conveyed at short notice: on average almost half of them (49.3%) are booked at most 10 minutes before their EPT. Even more surprising is the fact that a few requests (on average 0.6%) are booked a posteriori, that is, after their desired EPT. The data indicate that there is no strict policy on request notification and as a result, transport bookings reach the dispatcher very late. Clearly, this policy has a negative impact on the quality of the service provided by the transportation department. Not only are dispatching actions required to be very fast but it is also not always possible to have enough transportation resources available to provide service on time, in particular during the morning peak hours. This aspect will be made clear in Section 5.3.

Due to the limited monitoring means of the vehicle rides as mentioned above, the data do not include the actual pickup and delivery times at each visited location and so it is impossible to reconstruct the routes of the ambulances. Consequently, a comparison of the proposed algorithm with the manual dispatching rules cannot be carried out. Therefore, several operating scenarios and variants of the algorithm were tested, which are described in the next section.

5.2 Parameter settings

Only requests demanding service within the next $\Delta M = 8$ hours are processed by the heuristic. Following some preliminary experiments, the size of a neighborhood $N_p(i)$ of request i varies according to the number of outstanding requests. Thus, when the current vehicle routes include only a few requests, the number of neighbors to be considered is proportionally small. The opposite is also true. Moreover, a larger number of neighbors with respect to temporal proximity was considered compared to spatial proximity (i.e. $p_2 > p_1$). This choice was motivated by the fact that the hospital gives preference to reducing patient waiting times. Hence, we set $p_1 = \lfloor n/5 \rfloor$ and $p_2 = \lfloor n/4 \rfloor$, with n denoting the number of requests in the current vehicle routes that were not serviced yet.

In the TS heuristic, which is used in the improvement phase, the parameter controlling the intensity of the diversification takes the value $\lambda = 0.015$ as proposed by Cordeau and Laporte (2003b). Empirical experiments showed that this value is also appropriate in our case. The

term μ that guides the penalization of violated constraints was set to 0.2. This value avoids modifying too drastically the penalty factors α, β and γ for excess vehicle capacity, equipment and ride time, respectively. Every $\kappa = 5$ iterations, intra-route moves are performed to intensify the search in promising regions. Moreover, a move is declared tabu during $\theta = 7$ iterations. Finally, the insertion phase is run whenever a new request is booked (recall that the hospital data include the occurrence time of each request). The TS algorithm is then applied in an attempt to improve the initial routes. It stops either after $\ell_{max} = 15$ iterations or upon reaching a given time limit. The latter is set by the booking time of the next request. Preliminary experiments showed that a number of iterations larger than ℓ_{max} had little impact on the quality of the solutions obtained.

To assess the benefits of using the two-phase heuristic, experiments were also performed by applying simpler procedures. More specifically, the following approaches were tested:

P1: Only the insertion phase is performed.

P2: The full procedure (i.e. insertion followed by tabu search) is applied. In the inter-route step the choice of a route is based on the following hierarchical decision process: (i) the route with the highest lateness; (ii) if criterion (i) is not met, select the route with the largest excess ride time; (iii) if no routes satisfy criterion (ii), choose the route with the largest excess capacity; (iv) if criterion (iii) is not met, select the route with the largest excess equipment; (v) if criterion (iv) is not fulfilled, choose the route with the highest earliness. If none of the above criteria is met by any route then the route with worst total cost is selected.

P3: The full procedure is performed with a variant of the inter-route step so as not to allow violation of any constraints. Hence, a request is only moved from its current route to another route if this leads to a feasible solution with total cost lower than the best known so far. The same aspiration criterion as in the original TS algorithm is applied to revoke the tabu status of a move.

Upon selecting a route at an iteration of the improvement phase according to one of the rules described above, an outstanding request, without a pre-assigned status, is chosen for removal from its current vehicle route. The selection mechanism is driven by a hierarchical decision process. If a given criterion is not met by any request in the selected route then the next criterion is considered according to the following order:

1. choose the request with the highest lateness;

2. choose the request with the largest excess ride time;
3. choose the request with the largest excess capacity;
4. choose the request with the largest excess equipment;
5. choose the request with the highest earliness;
6. choose the request with the longest direct ride time.

In procedure P3 only criteria 1, 5 and 6 are used since route feasibility must be preserved at all times. Both the route and request selection mechanisms are motivated by the relative importance assigned by the hospital to each one of the three components in the objective function. The next section presents the numerical results obtained by applying P1–P3 to 20 real-life instances supplied by the hospital. For ease of maintenance and flexibility, the procedures were implemented using the programming framework of Microsoft Visual Basic for Applications (VBA). All tests were performed on a Pentium M with 1.8 GHz processor. The computing times obtained provide an upper bound on the CPU times that would be measured with a more sophisticated development platform and an object-oriented implementation language such as C^{++} or Java.

5.3 Numerical results

Table 7 summarizes the results obtained. Column 2 reports the total cost of the best solution identified at the end of each day. The percent deviation to that cost is shown in columns 3–5 for each procedure, that is, '% over best' $= (c(\text{Pi}) - \min_{1 \le i \le 3} c(\text{Pi})) \ / \ \min_{1 \le i \le 3} c(\text{Pi})$, with $c(\text{Pi})$ denoting the total cost obtained with procedure Pi, i $= 1, 2, 3$. Obviously, the best solution identified corresponds to a procedure with 0% deviation, which is highlighted with bold typeface. The last three columns in Table 7 present the total CPU time in minutes required by each procedure for a one-day simulation from 7:00 a.m. until 4:42 p.m.

The results show that higher quality solutions are identified during the improvement phase in procedures P2 and P3, thus demonstrating the usefulness of the TS algorithm. An improvement of at least 20% is obtained over the vehicle routes constructed during the insertion step (procedure P1). In some cases, e.g. days 4, 8 and 14, the impact of the post-optimization is remarkable. The best solutions are obtained by procedure P2 in 60% of the data sets. This indicates that a wider exploration of the search space allowing intermediate infeasible moves during the TS algorithm seems to be a good strategy. However, the superiority of P2 comes at the expense of significantly larger CPU time. Even though procedure P3 is only slightly worse

Day	Total cost of best solution	% over best			CPU time (min.)		
		P1	P2	P3	P1	P2	P3
1	690.5	37.9	**0.0**	1.0	3.4	69.1	23.3
2	657.5	48.9	**0.0**	3.1	2.5	42.9	10.4
3	586.0	24.8	10.8	**0.0**	2.1	28.1	11.0
4	754.1	136.3	**0.0**	14.6	4.3	69.5	44.4
5	715.1	57.9	6.2	**0.0**	4.1	51.4	26.0
6	632.7	23.1	**0.0**	0.7	1.8	32.2	12.6
7	645.6	25.3	**0.0**	4.9	3.6	34.1	15.7
8	859.7	99.6	**0.0**	0.4	4.8	80.7	44.9
9	534.0	22.6	1.1	**0.0**	1.6	26.5	7.0
10	637.4	44.5	4.1	**0.0**	2.3	57.6	15.2
11	740.9	75.7	**0.0**	8.6	2.4	40.3	16.7
12	647.5	34.2	**0.0**	1.5	2.6	33.8	15.8
13	652.2	31.9	**0.0**	4.4	2.6	29.8	17.1
14	794.6	146.3	5.4	**0.0**	3.4	76.6	28.5
15	648.4	37.7	7.5	**0.0**	2.8	40.4	16.7
16	526.1	34.8	**0.0**	10.5	2.1	29.0	10.0
17	689.8	19.8	**0.0**	8.1	2.4	50.4	12.4
18	645.9	45.6	**0.0**	4.1	2.0	33.7	12.8
19	831.6	33.4	7.1	**0.0**	3.3	50.5	18.4
20	717.6	24.1	1.5	**0.0**	2.2	30.0	14.5
Avg.	680.4	50.2	**2.2**	3.1	2.8	45.3	18.7

Table 7: Best solutions identified and CPU time

than P2 on average, solutions of different quality are generated by these two heuristics. In those instances in which procedure P3 is the best, the gap to procedure P2 is relatively large. The reverse also applies. For instance, on day 4, P3 is almost 15% more expensive than P2, while on day 3, P2 is about 11% worse than P3. The enforcement of feasibility in the TS algorithm seems to have a positive effect on the solution process in only 40% of the data sets. Since the CPU time required by procedure P2 is on average more than twice that of procedure P3, the latter may be an attractive alternative to P2 during those busy periods of the day when short response times are crucial. During the improvement step in procedures P2 and P3, the choice of a route from which a request is to be removed is guided by a hierarchical decision process as described in Section 5.2. We also conducted experiments with an alternative selection mechanism which first considers the route with the worst total cost in each iteration of the

TS algorithm. An outstanding request is then chosen from the considered route by following the same selection process as in procedures P2 and P3. This approach never generated the best solutions and is therefore omitted from the analysis of the results. Considering that the hospital gives higher priority to preventing lateness upon pickup or delivery of a patient, it is not surprising that procedures that make explicit use of this criterion, as is the case of P2 and P3, always outperform the former strategy.

The CPU time required for a one-day simulation, as reported in Table 7, proved to be adequate to process the large number of daily requests even though the three variants of the algorithm were run in a VBA environment which is a not a sophisticated platform.

Since the cost function of our problem combines three components with different relative importance, we have also evaluated their individual values. Table 8 shows the total duration of all routes as well as the total lateness and earliness at all locations in minutes. The last two terms refer to the actual values, that is, without the penalty factors. For each one of these components, the best value found is reported in Table 8 and the percent deviation in each procedure is computed. This allows us to assess the strengths and weaknesses of each heuristic.

Day	Total travel time				Total lateness				Total earliness			
	Best value	% over best			Best value	% over best			Best value	% over best		
	(min.)	P1	P2	P3	(min.)	P1	P2	P3	(min.)	P1	P2	P3
1	2056.0	17.8	**0.0**	3.0	85.3	304.9	7.1	**0.0**	55.2	55.3	71.4	**0.0**
2	2042.7	15.7	**0.0**	3.7	46.8	810.6	16.7	**0.0**	35.5	**0.0**	68.4	103.7
3	1826.9	8.5	10.5	**0.0**	55.6	264.8	**0.0**	3.4	17.1	332.5	198.4	**0.0**
4	2182.0	9.5	9.6	**0.0**	39.7	2784.0	**0.0**	684.9	25.0	170.2	106.2	**0.0**
5	2179.2	0.3	6.2	**0.0**	90.3	641.1	1.9	**0.0**	5.7	**0.0**	800.6	533.6
6	1959.9	9.8	3.6	**0.0**	20.0	888.2	**0.0**	233.1	45.8	31.1	25.8	**0.0**
7	1961.6	7.2	3.3	**0.0**	45.9	459.8	**0.0**	172.5	34.3	**0.0**	51.6	15.4
8	2581.5	4.6	**0.0**	2.5	100.7	1065.5	12.1	**0.0**	45.0	9.5	64.1	**0.0**
9	1552.2	13.1	10.6	**0.0**	22.0	800.1	**0.0**	280.2	39.8	**0.0**	42.8	127.0
10	1845.6	14.7	11.4	**0.0**	55.1	699.8	**0.0**	124.4	22.0	**0.0**	213.6	116.4
11	2369.3	9.5	**0.0**	4.6	37.6	1748.2	**0.0**	138.1	34.5	37.3	**0.0**	6.6
12	1908.4	3.2	5.3	**0.0**	54.0	671.1	**0.0**	123.2	44.0	**0.0**	41.0	40.7
13	2071.4	7.4	**0.0**	3.5	33.3	762.1	**0.0**	46.7	39.9	116.3	35.6	**0.0**
14	2279.4	21.2	4.3	**0.0**	168.9	620.5	9.4	**0.0**	47.3	9.3	42.4	**0.0**
15	1994.9	2.3	5.4	**0.0**	63.6	556.9	**0.0**	3.7	51.9	75.5	170.6	**0.0**
16	1710.2	3.2	0.5	**0.0**	9.9	2652.2	**0.0**	869.6	23.4	184.0	**0.0**	132.7
17	2138.5	6.2	2.2	**0.0**	40.6	412.5	**0.0**	277.0	49.3	47.1	**0.0**	28.2
18	1927.4	12.1	**0.0**	2.1	92.8	360.8	**0.0**	34.4	37.0	38.3	62.5	**0.0**
19	2533.0	1.7	4.4	**0.0**	99.7	290.4	37.7	**0.0**	53.9	**0.0**	6.9	10.0
20	2071.5	14.9	9.9	**0.0**	66.0	298.2	**0.0**	106.1	27.1	242.9	**0.0**	166.5
Avg.	2059.6	9.1	4.4	**1.0**	61.4	854.6	**4.2**	154.9	36.7	67.5	100.1	**64.0**

Table 8: Comparison of travel time, lateness and earliness

Procedure P3 tends to generate shorter routes than procedure P2. However, a reduction in ride time impacts negatively on waiting times for patients. Note that lateness is considerably larger in procedure P3 than in P2. The results clearly indicate that the main weakness of procedure P1 is its inability to service requests on time. The strength of procedures P2 and P3

lies in the inter-route and intra-route moves, which allow improving the overall service quality by reducing lateness significantly compared to P1. Regarding earliness, which is the least important cost component for the hospital, the results show a slight superiority of procedure P3. The cumulated earliness over the day and all vehicles is relatively small, being on average about 37 minutes. This indicates that vehicles do not frequently idle with patients on board, which is a positive feature. Total lateness is almost twice as large as total earliness. This is partly explained by the relatively late notification of new requests as shown in Table 6. Recall that on average almost half of the requests are booked less than 10 minutes prior to their earliest pickup time. Furthermore, the width of the time windows is rather small (15 minutes for the majority of requests). Hence, it is often impossible to arrive at an origin location within its time window. Due to the way time windows are specified, late pickup also implies late delivery.

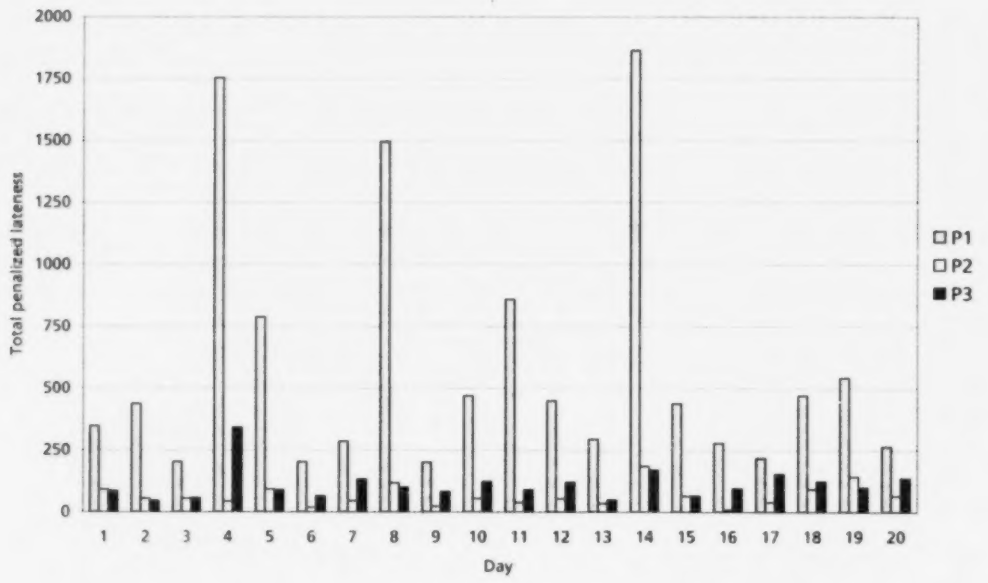

Figure 2: Total penalized lateness

Figure 2 depicts the value of $\sum_v \sum_{k=1}^{q} f(E^v_{n_k})$, that is, the daily total penalized lateness. Recall that if a vehicle arrives at a critical location after the desired latest time and lateness exceeds the tolerance specified by the priority of the corresponding request, then it is penalized by a factor of three. Procedures P2 and P3 tend to keep lateness below the tolerance value,

which clearly reduces patient inconvenience. Compared to P3, procedure P2 yields slightly better results in 70% of the data sets.

The analysis of Table 8 is complemented with the evaluation of the level of activity of the fleet of 11 ambulances. Columns 2-4 in Table 9 indicate for each procedure the total number of vehicles that serviced at least one request during each day. The remaining columns report the total time spent by the vehicles at the depot, which includes the work hours of unused ambulances as well as inactive time periods of used ambulances. The latter are originated by trips back to the depot that are made by vehicles when they do not carry any patients and their sojourn time is at least 10 minutes.

Day	# Vehicles used			Total inactivity time at depot			
	P1	P2	P3	Lowest value (min.)	% over lowest value P1	P2	P3
1	**8**	11	10	2453.3	**0.0**	14.6	13.7
2	**8**	10	**8**	2562.4	**0.0**	11.6	8.2
3	**8**	10	**8**	2892.9	0.4	**0.0**	7.8
4	**8**	10	8	2504.6	**0.0**	0.6	10.0
5	**7**	11	9	2597.8	6.7	**0.0**	5.7
6	**6**	10	9	2749.6	**0.0**	4.5	7.5
7	**8**	10	10	2823.9	**0.0**	2.1	4.8
8	**8**	10	10	2212.7	**0.0**	4.3	2.6
9	9	9	**8**	3167.2	**0.0**	0.7	4.8
10	**6**	10	7	2822.7	**0.0**	0.5	8.7
11	**9**	11	10	2320.7	**0.0**	10.2	5.5
12	**7**	8	8	2890.9	2.0	**0.0**	3.5
13	**8**	10	9	2651.4	**0.0**	7.0	4.7
14	**9**	11	11	2148.7	**0.0**	17.2	22.7
15	**6**	9	7	2719.0	4.1	**0.0**	8.3
16	**5**	8	7	3131.2	**0.0**	2.8	2.1
17	**8**	11	9	2617.5	**0.0**	4.2	5.5
18	**9**	10	10	2750.4	**0.0**	8.1	7.5
19	**9**	11	11	2259.7	3.2	**0.0**	4.9
20	**8**	10	8	2488.8	**0.0**	6.8	13.2
Avg.	7.7	10.0	8.9	2638.3	0.8	4.8	7.6

Table 9: Number of vehicles used and total time spent at the depot by the entire ambulance fleet

Although the minimization of the number of vehicles used is not explicitly considered in

the objective function, a side effect of the insertion heuristic (procedure P1) is to use as few vehicles as possible. Each time a new request is booked, a subset of spatial and temporal neighbors of its critical location is chosen among outstanding requests in the planned routes. Since the two vehicle depots are located relatively far away from most hospital service units, they are usually not attractive neighbors. Hence, the fleet minimization effect observed in procedure P1 is explained by the fact that new requests are often assigned to vehicles that are already riding on campus streets. As a result, the vehicle routes tend to be longer in procedure P1 than in the other two procedures, as confirmed by Table 8. Hence, each used vehicle has a much larger number of requests to service during certain periods of the day, making it more difficult to provide service on time. Consequently, the total lateness grows substantially as previously observed in Table 8. Furthermore, the total inactivity time increases in P1 since it includes the work periods of all unused vehicles. Procedures P2 and P3 tend to use more vehicles due to the exploration of possible benefits that can be obtained by modifying the routes constructed by procedure P1. In particular, an inter-route move allows to select a request that impacts negatively on the cost of its current route. Neighboring stops of this request are then determined. In some cases, the set of spatial neighbors includes vehicles still stationed at their home depot. The assignment of the request to such a vehicle involves a tradeoff between increasing the total travel time (because the depot may be far away from the origin and destination stops of the request), and reducing the total lateness by removing the request from its current route. In most cases, it pays off to perform this type of move. This strategy seems to be followed more often by procedure P2 than by procedure P3 as more vehicles are used by P2. From the patient's viewpoint, this strategy is beneficial since it yields a reduction of lateness. However, it also leads to unbalanced routes with respect to the total number of requests serviced by each vehicle and the corresponding total travel time. This aspect is more evident in procedure P2 than in P3. A detailed analysis of the vehicle routes has shown that only a few vehicles leave the depot for short periods of the day, while other vehicles are regularly busy.

The results obtained suggest that there are various opportunities for improving the transportation service. On the one hand, the ambulance fleet is underutilized. Sparing a single ambulance leads to substantial annual savings on maintenance costs. Moreover, staff reduction is not necessary since other tasks may be assigned to the freed ambulance personnel. On the other hand, it is not possible to avoid lateness completely with the current fleet of 11 ambulances. This is mainly due to the poor management of the transport booking system. Many

transportation requests originate from appointments for inpatients that are demanded by nursing wards to diagnostic and therapeutic facilities. Usually, a regular appointment is granted on the same day or on the following day. However, the transportation department is not notified as soon as an appointment is made. Nursing wards are responsible for communicating their transport needs and although this information is available in advance, it is conveyed to the transportation department very late. An integration of the transport booking system into the appointment system would greatly enhance the quality of the transportation services by reducing patient inconvenience and resource underutilization.

6 Conclusions

We have developed a two-phase heuristic procedure to solve a dynamic dial-a-ride problem arising in a hospital context. Our algorithm is designed to address hospital-specific features that considerably increase the complexity of the problem compared to the classic DARP. In particular, as time unfolds and new events occur, previously planned routes need to be modified in real-time. Therefore, response time is a crucial issue in a dynamic environment. In the first phase of our method, new requests are assigned and scheduled to existing vehicle routes by a simple insertion procedure. This approach makes use of spatial and temporal proximity measures to evaluate the level of attraction between pairs of requests, and thus identify promising insertion positions for a new request. The second phase of our algorithm attempts to improve the quality of the initial routes by means of a tabu search algorithm. Our study has shown that the post-optimization step yields significant improvements over the insertion method.

As is the case of most heuristic algorithms for large-scale pickup and delivery problems, it is difficult to state in quantitative terms the quality of the solutions identified. First, procedures for problems of similar size and type in a dynamic environment are not available and second, since patient transports are poorly managed in the German hospital where the study was conducted, we lack hand-made solutions. Nevertheless, based on the list of complaints received by the hospital, we believe that the solutions found by the best variant of our algorithm are superior to those produced in practice. In particular, waiting times for patients were reduced significantly using fewer vehicles. However, an analytical method alone is not sufficient to suppress the pitfalls of the current system. In a hospital context, improvements can only be achieved by reorganizing work processes and increasing awareness among hospital staff about the importance of efficient logistics practices. Similar to the study reported by Banerjea-Brodeur et al. (1998),

we also experienced that common sense reasoning already yields better planning of operations. This is true with respect to our recommendation for nursing wards and hospital service units to book new requests for transportation as soon as they become known. As shown before, the current practice is responsible for creating considerable patient inconvenience.

The hospital is considering introducing a wireless local-area network (WLAN) and equipping the ambulances with hand-held computers. This would greatly enhance the communication between the vehicles and the dispatch office. Since communication would take place at any time and not just at service locations, current research is focusing on extending the algorithm to include the possibility of diverting a vehicle away from its immediate destination to service a new request in the vicinity of the vehicle's current position. Furthermore, additional stochastic elements such as service and travel times should be considered. The latter are influenced by waiting time during congested periods of elevator use. Hence, recourse actions should be developed to cope with unexpected events such as high elevator congestion.

Appendix - Derivation of temporal proximity measures

Let τ_{ij}, resp. τ_{ji}, denote the temporal proximity between requests i and j when the critical location of i is serviced before, resp. after, the critical location of j. The proximity measure is only calculated when the two requests are compatible. Observe that a zero proximity means that there is one single way of scheduling requests i and j without violating the time windows of their critical locations. Moreover, the larger the proximity, the larger the number of alternative ways of scheduling both requests on time.

In what follows we denote by $[a_i, b_i]$, resp. $[a_j, b_j]$, the time window at the critical location of request i, resp. of request j. Furthermore, we assume that the direct ride times satisfy the triangle inequality.

The *pickup* locations of requests i and j are critical

To calculate τ_{ij} we consider all paths linking $o(i)$ and $o(j)$ directly, that is, paths (2) and (3). If $a_i + DRT_{o(i),o(j)} > b_j$ then the sequence $o(i) \rightarrow o(j)$ is clearly incompatible, otherwise there is at least one departure time from $o(i)$ that ensures a timely arrival at $o(j)$. Hence, if condition $a_i + DRT_{o(i),o(j)} \leq b_j$ holds, the temporal proximity is given by

$$\tau_{ij} = \min\{b_j, b_i + DRT_{o(i),o(j)}\} - \big(a_i + DRT_{o(i),o(j)}\big) + \max\{0, a_j - b_i - DRT_{o(i),o(j)}\}. \quad (9)$$

The first term measures the latest feasible arrival time at $o(j)$, while the second term refers to the earliest feasible arrival time at $o(j)$. The last term measures the minimum waiting time at $o(j)$ in the case of an early arrival at that location, which can occur when the time windows are far apart. The calculation of τ_{ji} is based on a similar analysis applied to paths (4) and (5), and yields (9) by interchanging i with j.

The *delivery* location of request i and the *pickup* location of request j are critical

A close examination of path (6), where $d(i)$ is the predecessor of $o(j)$, shows that the non-critical stop $o(i)$ is visited immediately before $d(i)$. This means that it is always possible to arrive on time at $d(i)$ from $o(i)$, and so location $o(i)$ can be disregarded from the calculation of τ_{ij}. A further implication of this observation is that the time window $[a_i, b_i]$ of service point $d(i)$ has the same status of a pickup location. As a result, τ_{ij} is also given by (9).

To calculate τ_{ji} we analyze sequence $o(i) \rightarrow o(j) \rightarrow d(i)$ in path (2). Although the pickup point $o(i)$ is not critical, departing later than LPT_i will certainly violate the time window of its drop-off point $d(i)$. Moreover, an incompatibility between the time windows of the critical stops $o(j)$ and $d(i)$ may also be caused by the impossibility of servicing the pickup point $o(j)$ within its time window. This occurs when $a_{i'} + DRT_{o(i),o(j)} > b_j$ with i' representing $o(i)$ and $[a_{i'}, b_{i'}]$ denoting its time window. The reverse inequality, however, does not necessarily guarantee the desired compatibility. Let $\widetilde{a}_j$, resp. $\widetilde{b}_j$, be the earliest, resp. latest, departure time from $o(j)$. Observe that if inequality $\widetilde{a}_j + DRT_{o(j),d(i)} > b_i$ holds then locations $o(j)$ and $d(i)$ are certainly incompatible. Summarizing, if conditions $a_{i'} + DRT_{o(i),o(j)} \leq b_j$ and $\widetilde{a}_j + DRT_{o(j),d(i)} \leq b_i$ hold then the temporal proximity is measured by

$$\tau_{ji} = \min\{b_i, \widetilde{b}_j + DRT_{o(j),d(i)}\} - \left(\widetilde{a}_j + DRT_{o(j),d(i)}\right). \tag{10}$$

Observe that $\widetilde{a}_j = \min\{b_j, \max\{a_j, a_{i'} + DRT_{o(i),o(j)}\}\}$. The latest feasible departure time from $o(i)$ influences the value of $\widetilde{b}_j$ and is given by $\widetilde{b}_{i'} = \max\{a_{i'}, b_j - DRT_{o(i),o(j)}\}$. Hence, $\widetilde{b}_j = \min\{b_j, \max\{a_j, \widetilde{b}_{i'} + DRT_{o(i),o(j)}\}\}$.

The *pickup* location of request i and the *delivery* location of request j are critical

When location $o(i)$ is visited immediately before location $d(j)$, the calculation of the temporal proximity relies on the analysis of path (5). A close examination indicates that only the sequence

$o(j) \rightarrow o(i) \rightarrow d(j)$ is of interest, and so the derivation of τ_{ij} follows the same steps as in the previously studied case. Thus, expression (10) yields τ_{ij} by simply interchanging i with j everywhere. To determine τ_{ji}, path (7) is analyzed. Following the same reasoning as in the calculation of τ_{ij} when locations $o(j)$ and $d(i)$ are critical, we conclude that only the sequence $d(j) \rightarrow o(i)$ is relevant to obtain now τ_{ji}. Consequently, the temporal proximity is measured by (9) with i interchanged with j everywhere.

The *delivery* locations of requests i and j are critical

Let the time window of the non-critical stop $o(i)$, resp. $o(j)$, be denoted by $[a_{i'}, b_{i'}]$, resp. $[a_{j'}, b_{j'}]$. The temporal closeness τ_{ij} depends on two different ways of linking stops $d(i)$ and $d(j)$ directly, and is therefore given by $\max\{\tau_{ij}^1, \tau_{ij}^2\}$ with τ_{ij}^1 representing the proximity according to path (2), and τ_{ij}^2 the proximity in path (4).

In what follows we represent the earliest, resp. latest, departure time from location k by $\widetilde{a}_k^1$, resp. $\widetilde{b}_k^1$, in path (2) with $k = i', j', i$. Compatibility between locations $d(i)$ and $d(j)$ is ensured when they can be serviced without violating the upper bounds of their time windows. This implies that if $\widetilde{a}_{j'}^1 + DRT_{o(j),d(i)} \leq b_i$ and $\widetilde{a}_i^1 + DRT_{d(i),d(j)} \leq b_j$ then the proximity is given by

$$\tau_{ij}^1 = \min\{b_j, \widetilde{b}_i^1 + DRT_{d(i),d(j)}\} - \left(\widetilde{a}_i^1 + DRT_{d(i),d(j)}\right). \tag{11}$$

Clearly, $\widetilde{a}_i^1$ depends on the earliest departure time from stop $o(j)$, which in turn depends on the departure time from stop $o(i)$. Hence, $\widetilde{a}_i^1 = \min\{b_i, \widetilde{a}_{j'}^1 + DRT_{o(j),d(i)}\}$ and $\widetilde{a}_{j'}^1 = \max\{a_{j'}, a_{i'} + DRT_{o(i),o(j)}\}$. Since stops $o(i)$ and $o(j)$ are both non-critical, in principle we can leave these locations after their latest pickup times but certainly not before their earliest pickup times. It follows that $\widetilde{b}_{i'}^1 = \max\{a_{i'}, \widetilde{b}_{j'}^1 - DRT_{o(i),o(j)}\}$ and $\widetilde{b}_{j'}^1 = \max\{a_{j'}, \overline{b}_i - DRT_{o(j),d(i)}\}$ with $\overline{b}_i = \min\{b_i, b_j - DRT_{d(i),d(j)}\}$. Thus, $\widetilde{b}_i^1 = \min\{b_i, \widetilde{b}_{i'}^1 + DRT_{o(i),o(j)} + DRT_{o(j),d(i)}\}$. Finally, the proximity τ_{ij}^2 associated with path (4) is determined by following the same steps as above and interchanging $o(i)$ with $o(j)$. Regarding the temporal proximity when location $d(j)$ is serviced before location $d(i)$, i.e. τ_{ji}, it depends on paths (3) and (5), and is derived by using the same reasoning as for τ_{ij} but reversing the order of i and j, and i' and j'.

Temporal closeness and pre-assigned requests

Since request j is serviced by a particular vehicle, the first stop in the relevant paths (4), (5) and (7) must be identified. Note that this stop is the immediate destination of the vehicle

which may not coincide neither with the pickup point $o(j)$ nor with its drop-off point $d(j)$. Furthermore, the time window of the first stop is replaced by the estimated time of service by the vehicle. The temporal proximity is then calculated according to one of the cases derived above depending on the combination of critical locations.

Acknowledgements

This research was partially funded by the National Sciences and Engineering Research Council of Canada under grant 39682-05, and by the German Academic Exchange Service under project PPP Kanada D/05/50480 "Optimization of Transports in Hospitals". This support is gratefully acknowledged. The authors also wish to thank the German hospital for providing data, technical expertise and valuable assistance throughout this research project.

References

Aldaihani, M., M. Dessouky. 2003. Hybrid scheduling methods for paratransit operations. *Computers & Industrial Engineering* **45** 75–96.

Attanasio, A., J.-F. Cordeau, G. Ghiani, G. Laporte. 2004. Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. *Parallel Computing* **30** 377–387.

Banerjea-Brodeur, M., J.-F. Cordeau, G. Laporte, A. Lasry. 1998. Scheduling linen deliveries in a large hospital. *Journal of the Operational Research Society* **49** 777–780.

Baugh, J., G. Krishna, R. Kakivaya, J. Stone. 1998. Intractability of the dial-a-ride problem and a multiobjective solution using simulated annealing. *Engineering Optimization* **30** 91–123.

Bent, R., P. Van Hentenryck. 2006. A two-stage hybrid algorithm for pickup and delivery vehicle routing problems with time windows. *Computers & Operations Research* **33** 875–893.

Bergvinsdottir, K.B., J. Larsen, R. Jørgensen. 2004. Solving the dial-a-ride problem using genetic algorithms. *Journal of the Operational Research Society,* forthcoming.

Bodin, L.D., T. Sexton. 1986. The multi-vehicle subscriber dial-a-ride problem. *TIMS Studies in Management Science* **22** 73–86.

Borndörfer, R., M. Grötschel, F. Klostermeister, C. Küttner. 1997. Telebus Berlin: Vehicle scheduling in a dial-a-ride system. Technical Report SC 97-23, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Germany. Available online at www.zib.de/PaperWeb/abstracts/SC-97-23.

Cordeau, J.-F. 2006. A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research* **54** 573–586.

Cordeau, J.-F., G. Laporte. 2003a. The dial-a-ride problem (DARP): Variants, modeling issues and algorithms. *4OR: A Quarterly Journal of Operations Research* **1** 89–101.

Cordeau, J.-F., G. Laporte. 2003b. A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research Part B* **37** 579–594.

Coslovich, L., R. Pesenti, W. Ukovich. 2006. A two-phase insertion technique of unexpected customers for a dynamic dial-a-ride problem. *European Journal of Operational Research* **175** 1605–1615.

Desrosiers, J., Y. Dumas, F. Soumis. 1986. A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. *American Journal of Mathematical and Management Sciences* **6** 301–325.

Desrosiers, J., Y. Dumas, F. Soumis. 1988. The multiple vehicle dial-a-ride problem. J. Daduna, A. Wren, eds., *Computer-Aided Transit Scheduling, Lecture Notes in Economics and Mathematical Systems*, vol. 308. Springer, Berlin, 15–27.

Diana, M., M. Dessouky. 2004. A new regret insertion heuristic for solving large-scale dial-a-ride problems with time windows. *Transportation Research Part B* **38** 539–557.

Dumas, Y., J. Desrosiers, F. Soumis. 1989. Large scale multi-vehicle dial-a-ride problems. Les Cahiers du GERAD G-89-30, École des Hautes Études Commerciales, Montréal, Canada.

Fu, L. 2002. Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. *Transportation Research Part B* **36** 485–506.

Gendreau, M., F. Guertin, J.-Y. Potvin, R. Séguin. 1998a. Neighbourhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. *Transportation Research Part C*, forthcoming.

Gendreau, M., A. Hertz, G. Laporte, M. Stan. 1998b. A generalized insertion heuristic for the traveling salesman problem with time windows. *Operations Research* **43** 330–335.

Jaw, J., A.R. Odoni, H.N. Psaraftis, N.H.M. Wilson. 1986. A heuristic algorithm for the multi-vehicle advance-request dial-a-ride problem with time windows. *Transportation Research Part B* **20** 243–257.

Landry, S., R. Philippe. 2004. How logistics can service healthcare. *Supply Chain Forum* **5** 24–30.

Madsen, O.B.G., H.F. Ravn, J.M. Rygaard. 1995. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research* **60** 193–208.

Melachrinoudis, E., A. Ilhan, H. Min. 2007. A dial-a-ride problem for client transportation in a health-care organization. *Computers & Operations Research* **34** 742–759.

Nanry, W.P., J.W. Barnes. 2000. Solving the pickup and delivery problem with time windows using reactive tabu search. *Transportation Research Part B* **34** 107–121.

Psaraftis, H.N. 1980. A dynamic programming solution to the single-vehicle, many-to-many immediate request dial-a-ride problem. *Transportation Science* **14** 130–154.

Psaraftis, H.N. 1983. An exact algorithm for the single-vehicle, many-to-many dial-a-ride problem with time windows. *Transportation Science* **17** 351–357.

Ropke, S., J.-F. Cordeau, G. Laporte. 2006. Models and branch-and-cut algorithms for pickup and delivery problems. *Networks*, forthcoming.

Ropke, S., D. Pisinger. 2006. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation Science* **40** 455–472.

Savelsbergh, M.W.P. 1985. Local search in routing problems with time windows. *Annals of Operations Research* **4** 285–305.

Sexton, T., Y.-M. Choi. 1986. Pickup and delivery of partial loads with soft time windows. *American Journal of Mathematical and Management Sciences* **6** 369–398.

Taillard, É.D. 1993. Parallel iterative search methods for vehicle routing problems. *Networks* **23** 661–673.

Toth, P., D. Vigo. 1996. Fast local search algorithms for the handicapped persons transportation problem. I. H. Osman, J. P. Kelly, eds., *Meta-heuristics: Theory and Applications*. Kluwer, Boston, 677–690.

Toth, P., D. Vigo. 1997. Heuristic algorithms for the handicapped persons transportation problem. *Transportation Science* **31** 60–71.

Wilson, N.H.M., N. Colvin. 1977. Computer control of the Rochester dial-a-ride system. Technical Report R-77-31, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Wilson, N.H.M., J.M. Sussman, H.K. Wong, B.T. Higonnet. 1971. Scheduling algorithms for dial-a-ride systems. Technical Report TR-70-13, Urban Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Wolfler Calvo, R., A. Colorni. 2006. An effective and fast heuristic for the dial-a-ride problem. *4OR: A Quarterly Journal of Operations Research*, forthcoming.